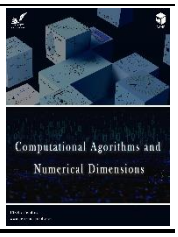
Computational Algorithms and Numerical Dimensions

www.journal-cand.com

Com. Alg. Num. Dim Vol. 2, No. 3 (2023) 124–135.



Paper Type: Original Article

Enhancing Water Pump Failure Prediction Using Machine Learning: A Focus on Less-Explored Variables

Soheil Azizi Borojerdi[1,*], Goran Cirovic[2]

[1] Faculty of Accounting and Management, Allameh Tabatabae University, Tehran, Iran; aziziborojerdi@gmail.com.

[2] Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovica-6, 21000 Novi Sad, Serbia; goran.cirovic@uns.ac.rs.

Citation:

Azizi Borojerdi, S., & Cirovic, G. (2023). Enhancing water pump failure prediction using machine learning: a focus on less-explored variables. *Computational algorithms and numerical dimensions, 2*(3), 124-135.

Received: 16/03/2023 Reviewed: 17/04/2023 Revised: 09/05/2023 Accepted: 01/06/2023

Abstract

In recent years, there has been a surge in research exploring the potential of Machine Learning (ML) for predicting water pump failures. While some studies have focused on supervised approaches, others have delved into unsupervised methods. However, the challenge lies in identifying the key variables crucial for accurate failure predictions. This study bridges this gap by consulting domain experts to discern essential variables, including water catchment area level, water quality index, lubrication frequency, water reservoir temperature, operating time, and power interruptions count. Employing supervised ML methods, specifically multiple regression and decision tree cart, the research aims to enhance the precision of failure predictions, shedding light on less-explored variables that play a significant role in pump failure.

Keywords: Machine learning, Water pump failure prediction, Multi-variable regression, Decision tree CART.

1 | Introduction

Accounting for reliability indicators has the potential to cut down operating costs significantly. Enhancements in reliability should be grounded in the accumulated operational experience of the National Assembly. Consequently, in 2008, TIIAME established a system for gathering and processing statistical data on the reliability of large NSs in operation and their components. The selection of information is intended to transition gradually from a passive state (as it currently exists) to an active state (as needed) [1]. The collection methodology is designed to concentrate relevant data in a format accessible to operational personnel. Analyzing these statistics enables the development of appropriate measures, leading to a substantial reduction in expensive research. The industry's focus on reliability is equally applicable to hydroelectric facilities. Notably, there has been an increased interest in recent times regarding the reliability of hydraulic structures. While there is a conservative approach to design using tried and tested methods, it should not impede the adoption of new techniques for analyzing reliability indicators, which include measures for technical diagnostics and maintainability [2].

Corresponding Author: aziziborojerdi@gmail.com

https://doi.org/10.22105/cand.2024.437591.1088

In recent times, there has been a surge of interest in the field of data mining across various domains. Data mining involves the process of uncovering valuable knowledge from extensive datasets. One of the primary tasks within data mining is addressing classification problems for diverse applications. These problems involve predicting categorical labels, which are the focal point of various domains, such as failure prediction in pumping stations [3]-[5]. To tackle classification tasks, previous experiences are initially gathered and represented. Subsequently, a predictive classifier is trained to describe the predefined set of data concepts. When faced with a new problem, the classifier can recommend corresponding solutions based on its training. Since the class label of each sample is required during the training process, data classification falls under the category of supervised learning. Evaluating the predictive accuracy of the classifier is often crucial when determining its applicability in a specific domain [6]-[11].

Several techniques can be employed in the data mining process. An effort was made by [12] to identify the top 10 data mining algorithms during the IEEE International Conference on Data Mining (ICDM) in December 2006. It involved inviting winners of the IEEE ICDM research contribution award and ACM KDD innovation award to nominate up to 10 algorithms in data mining. Each nomination was verified for citations on Google Scholar, and those with at least 50 citations were retained. The final top 10 algorithms were determined through an open vote by all attendees of ICDM. Among these top algorithms, Support Vector Machine (SVM) [13], Classification And Regression Tree (CART), C4.5 [14], [15], K-Nearest Neighbors (KNN) [16], and naïve Bayes [17] are commonly used techniques for classification mining.

In recent years, numerous articles have been presented on the application of Machine Learning (ML) in predicting failures for water pump systems. Some of these studies delve into supervised methods [18]-[20], while others explore unsupervised approaches [21]. However, the variables required for predicting failures have consistently posed challenges. This paper addresses this gap by consulting experts to identify variables crucial for predicting failures, including water catchment area level, water quality index, lubrication frequency, water reservoir temperature, operating time, and power interruptions count. These variables, which have received less attention in previous research, are examined using supervised methods, specifically multiple-variable regression and decision tree cart. The continuation of this paper involves a literature review encompassing data mining and various methodologies. The research method and dataset type are then discussed, followed by a comprehensive exploration of two approaches: multiple-variable regression and decision tree cart.

2 | Literature Review

2.1 | Machine Learning Overview

ML involves constructing an inductive model that autonomously learns from a limited dataset without requiring specialized intervention. This learning process involves identifying an underlying set of structures or patterns that prove valuable for comprehending relationships in data, even when it deviates from the original learning dataset. In the ML model taxonomy (*Fig. 1*), supervised learning predicts an output variable using labeled input data. In contrast, unsupervised learning makes inferences from unlabeled input data, as seen in clustering algorithms and recommender systems, among others. In supervised learning, distinctions are made between models predicting numeric variables (regression) and categorical variables (classifiers). Learning within models entails adjusting the model's parameters to a specific dataset and continually refining them through multiple passes of the data until a predefined function is minimized [22].

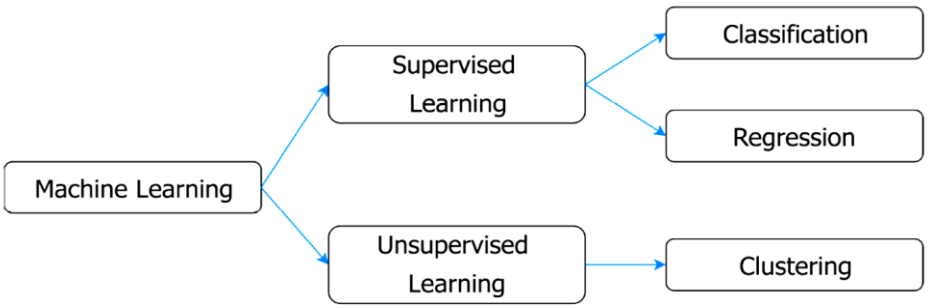


Fig. 1. The categorization of ML models.

2.2 | Regression

The technique of "Multiple Linear Regression" is one of the common methods in multivariate analysis. According to regression analysis, a linear relationship is established between the "response variable" and one or more "explanatory variables." The response variable is usually referred to as the "dependent variable," and the explanatory variables are called "independent variables." In the multiple linear regression method, the parameters of a linear model are estimated. Essentially, linear regression represents a linear relationship in terms of the model parameters. If we have n observations of a p-dimensional explanatory variable X and assume a linear relationship with the response variable y, we use the linear regression *Model (1)* [23].

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad (i = 1, \dots, n). \tag{1}$$

Considering that the explanatory variable X has p dimensions, its value in each dimension is replaced by a one-dimensional independent variable. Note that the subscript or index i represents the observation number. At the end of the linear relationship, there is also the error term ε, which represents the regression model's error.

2.3 | Decision Tree

A decision tree is one of the ML techniques used to organize decision-making algorithms. An English decision tree algorithm is employed for classifying features in a dataset using a cost function. This algorithm grows before optimization and pruning branches that do not contain valuable information, encompassing features unrelated to the problem. Therefore, pruning operations are performed to eliminate these extra branches [24].

In this algorithm, parameters such as the depth of the decision tree are adjustable to prevent overfitting or excessive complexity of the tree as much as possible. Various types of decision trees in ML are utilized for classification problems based on the characteristics of the training data. This technique can also be used for regression problems or predictions of response variable values for unseen data. The main advantage of employing ML algorithms in problems is their simplicity, as they make the decision-making process easily understandable. However, with an increase in the number of tree branches, understanding and applying the algorithm may become challenging due to excessive complexity. Therefore, pruning the tree becomes essential in such cases. Generally, a decision tree is considered an algorithm to calculate the potential success of diverse decision-making sequences in achieving a specific goal. Steps for implementing a decision tree [24]:

I. The first step is creating the top node, i.e., the root node, which includes the entire dataset.
II. Using a feature selection criterion method, the most important feature in the data is chosen.
III. Dividing the root node into child nodes that contain critical values for the most important features.
IV. Creating decision tree nodes constructed from the most important features.

V. Using the subsets created from the dataset in the third step, new decision trees are recursively created. This process continues until nodes can no longer be further divided, and terminal nodes, known as leaf nodes, are obtained.

In summary, the decision tree algorithm starts from the root node, selects the most important feature at each step, divides the dataset based on critical values for that feature, creates decision tree nodes, and recursively repeats the process until terminal nodes are formed.

3 | Methodology

In this study, to predict the number of failures per month for water booster pumps in the urban water treatment system, first, through interviews with experts in this field and reviewing the literature, factors and variables affecting the failure of this equipment were extracted. Then, with the permission of the Water and Wastewater Department, the required data was extracted from the reported data sets by the treatment plant. The data used in this research includes the values of variables collected by the pump maintenance officer over 168 weeks.

Given that the beginning of any work and operation initially involves a set of preliminaries and preparations, data mining is no exception to this, always requiring preparations and initial processing, known as the preprocessing stage. Preprocessing plays a crucial role in the data processing process and, subsequently, in the obtained results. In this study, after receiving the data, preprocessing operations were performed on the data, and outlier data was extracted from the data set. After removing incompatible data and replacing missing values with median values, the data was prepared for processing.

In the next stage, regression and decision tree methods were used to create a model for predicting the number of failures. Since ML models and data mining require the division of data into two groups, training and testing sets (and considering that the number of data points after preprocessing is 164, which is not a large number for ML models), we considered 50% of the data as test data and the remaining 50% as training data. In the next step, the regression method was implemented on the data, the basic assumptions of the regression model were examined, and finally, the results were investigated.

Continuing to examine the data further, decision tree methodology and the C5.0 algorithm were used to extract rules for pump failures. However, before implementing the algorithm, the target variable data was transformed into categorical values to allow the C5.0 algorithm to extract rules more accurately. In the end, the results obtained from both models were compared with each other. The research execution stages are shown in *Fig. 2*.

Table 1. Descriptive variables.

Variable	Symbol	Role	Measurement Unit
Water catchment area level	W. L	INPUT	Decimeter
Water quality index	WQI	INPUT	-
Lubrication frequency	LUB	INPUT	-
Water reservoir temperature	TEMP	INPUT	Celsius
Operating time	OP. TIME	INPUT	Hour (HOUR)
Power interruptions count	P.OUT	INPUT	-
Number of failures	NUM OF FAIL	OUTPUT	-

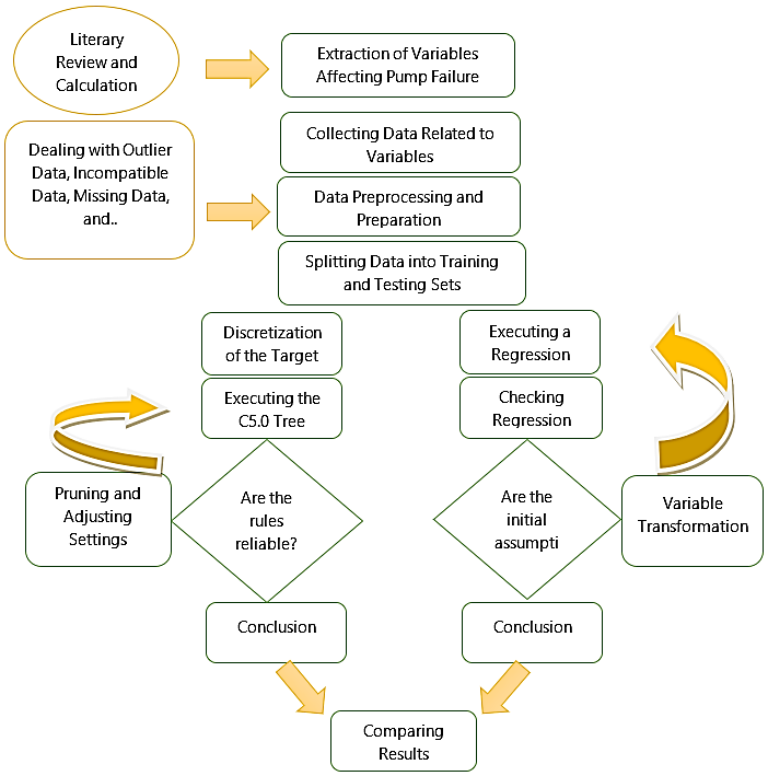


Fig. 2. Research execution stages.

4 | Dataset and Research Variables

The data for this study has been extracted from the Tehran purification plant in the city of Tehran. Following interviews with experts, the descriptive variables under investigation in this study are outlined in the *Table 1*.

The descriptive statistics related to the collected data for variables are presented in *Figs. 3-6*, which is the output of IBM SPSS MODELER software.

W.L	Continuous	28.680	49.860	38.555	3.965	-0.150	--	1
WQI	Continuous	27.970	83.200	54.465	11.036	0.002	--	1
P.OUT	Continuous	1.000	15.000	7.799	4.505	0.082	--	1
LUB	Continuous	0.000	10.000	4.494	2.923	0.217	--	1
OP. TIME	Continuous	85.000	455.000	233.329	90.745	0.540	--	1
TEMP	Continuous	9.000	44.000	28.970	6.196	0.015	--	1
NUM OF FAIL...	Continuous	0.000	15.000	4.561	3.170	0.735	--	1

Fig. 3. Descriptive statistics of the problem data.

5 | Multiple Linear Regression Analysis

After performing multiple linear regression on the variables, the predictor importance plot was obtained in *Fig. 3*. Typically, in modeling, efforts are made to focus on the most important predictor variables and either exclude or ignore those with less importance. The predictor importance plot, by indicating the relative importance of each predictor in estimating the model, helps you make these decisions. According to the results, the most important variable in pump failures is the power outage. In the next step, the temperature of the reservoir has a significant impact on the number of failures. The variable representing the operating time also has relatively high importance on the occurrence of failures. Lubrication frequency, water quality index, and water catchment area level variables are ranked third to sixth in importance, respectively. It is worth noting that while the water quality index and reservoir water level variables have less importance compared to other variables, based on expert opinions and the level of importance estimated from the model, their importance is not so low that they should be excluded from the model.

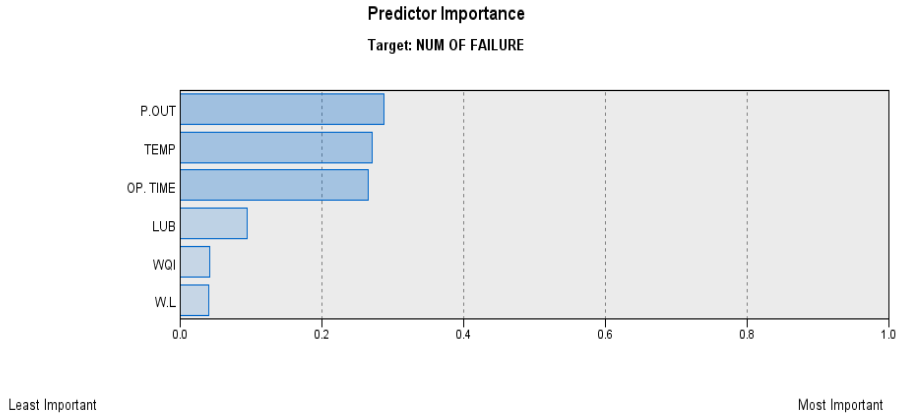


Fig. 4. Predictor importance plot.

The analysis of variance table (*Table 2*) for the regression model indicates that the obtained regression model is statistically significant, with a P-value less than 0.05. The F-statistic (F = 25.41) further suggests a high accuracy of the regression model. In *Table 3*, the regression coefficients of standardized variables show that the water reservoir level, continuous operating time, power outage frequency, and temperature positively influence the failure rate. Conversely, the water quality index and lubrication frequency negatively impact the failure rate. All descriptive variables are statistically significant, as indicated by the p-values less than 0.05. Overall, the regression model provides valuable insights into the relationship between various factors and the number of failures per month.

$$\text{Num OF Failure} = -1.896 + 0.032 * \text{W. L} - 0.051 * \text{WQI} + 0.172 * \text{P. OUT} - 0.136 * \text{LUB} + 0.100 * \text{OP. TIME} + 0.146 * \text{TEMP}.$$

Table 2. The analysis of variance.

ANOVA Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	417.312	6	69.552	25.421	.000[b]
	Residual	196.992	72	2.736		
	Total	614.304	78			

b. Predictors: (Constant), TEMP, LUB, WQI, P.OUT, W.L, OP. TIME

Table 3. The regression coefficients of the variables.

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-1.896	1.540		-1.232	.222		
	W.L	.032	.035	.071	.913	.047	.849	1.178
	WQI	-.051	.002	-.087	-.392	.046	.972	1.029
	P.OUT	.172	.047	.246	3.694	.000	.392	2.548
	LUB	-.136	.060	-.136	-2.263	.027	.484	2.067
	OP. TIME	.100	.002	.280	4.533	.000	.457	2.187
	TEMP	.146	.023	.426	6.387	.000	.391	2.557

Fig. 4 displays the output of the analysis node from the Modeler software. Based on the obtained results, the mean error in the training data is 0.00, and in the test data, it is 0.124, both very close to zero. The linear correlation between the actual target values (y) and predicted values (ŷ) in the training data is 0.835, and in the test data, it is 0.788, both very close to one. *Fig. 5* illustrates the plot of predicted values against actual target values. As the figure indicates, there is a high correlation between the target and predicted values.

Results for output field NUM OF FAILURE

Comparing $E-NUM OF FAILURE with NUM OF FAILURE

'Partition'	1_Training	2_Testing
Minimum Error	-2.001	-2.767
Maximum Error	2.34	5.519
Mean Error	0.0	0.124
Mean Absolute Error	0.786	0.89
Standard Deviation	0.994	1.226
Linear Correlation	0.835	0.788
Occurrences	79	85

Fig. 5. Output of the analysis node in IBM SPSS modeler.

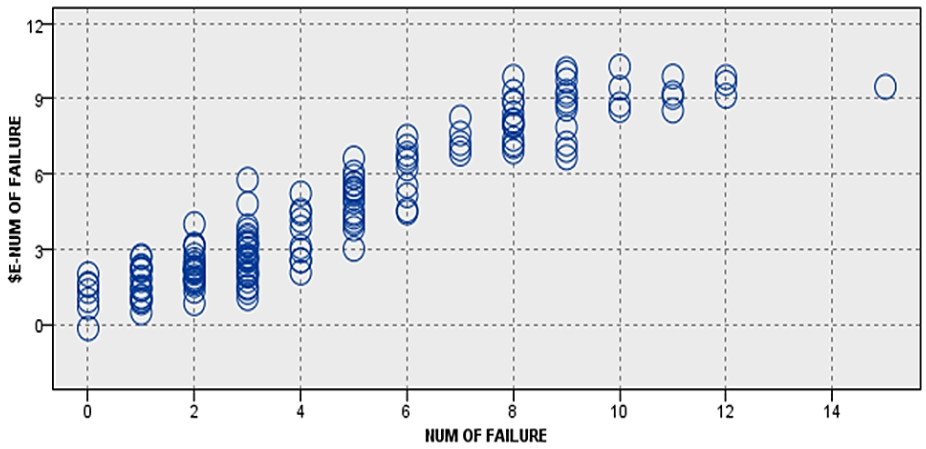


Fig. 6. Predicted values vs. actual target values.

To examine the lack of autocorrelation in the data, one can plot the order of data points (data number) against the errors. If no specific trend or pattern is observed, it can be assumed that the data points are not autocorrelated. *Fig. 6* illustrates the plot of the order of data points against error values. As depicted in *Fig. 6*, there is no discernible trend in error values with the progression of data points (x-axis).

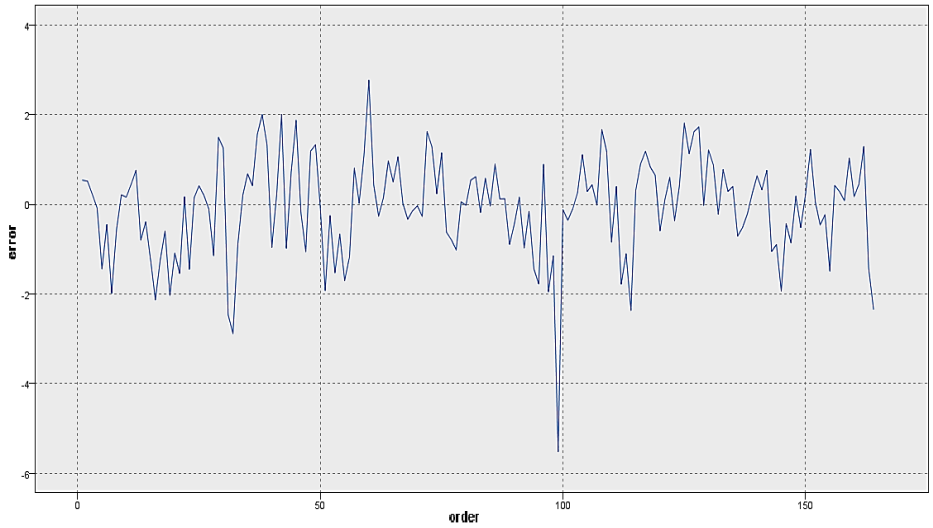


Fig. 7. The plot of error values against the order of data points.

6 | Using the C5 Decision Tree Algorithm for Predicting the Number of Failures

In this section, we aim to estimate the number of failures in water treatment pumps and examine the effects of variables, extracting rules for prediction using the C5 decision tree algorithm. Since the C5 algorithm is one of the supervised algorithms, where the target field, or the number of failures in this case, needs to be a categorical variable with categorical data, we first need to convert the values of the target field or the variable representing the number of failures into a categorical variable. To do this, after consulting with experts and observing the histograms of the data in the target field, four categories have been considered for this variable, as shown in *Table 4*. *Fig. 7* illustrates the histogram of the target field data and its separating bands.

Table 4. Target field classification.

Category	Description of Classification	Formula
1	Number of failures less than 3	'NUM OF FAILURE' < 3
2	Number of failures greater than or equal to 3 and less than 5	'NUM OF FAILURE' >= 3 and < 5
3	Number of failures greater than or equal to 5 and less than 8	'NUM OF FAILURE' >= 5 and < 8
4	Number of failures greater than or equal to 8	'NUM OF FAILURE' >= 8

In this phase, after transforming the target field into categorical data, the dataset is divided into training and testing sets, each comprising 50% of the data. The C5 algorithm is then applied using Modeler software with specific settings for tree pruning and leaf minimum records. The executed regression tree, as depicted in *Figs. 4-13*, reveals that temperature is the most influential factor in causing failures, followed by operation time and oiling frequency. The tree, benefiting from pruning, incorporates water quality and river water level variables in rule extraction. The resulting rules are based on the values of the number of oiling instances, operation time, and temperature. The Predictor Importance table in *Fig. 8* provides insights into the significance of each predictor in the decision tree.

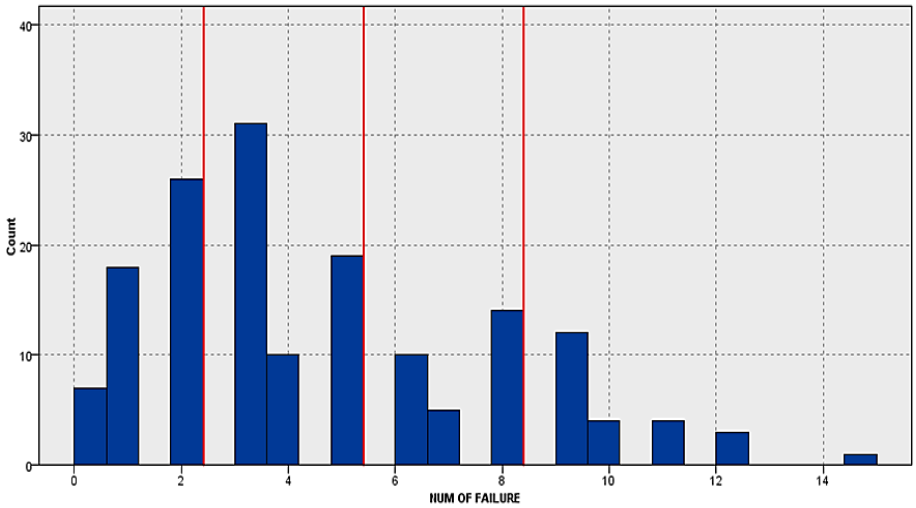


Fig. 8. The division of data is done to transform the target field data into categorical data.

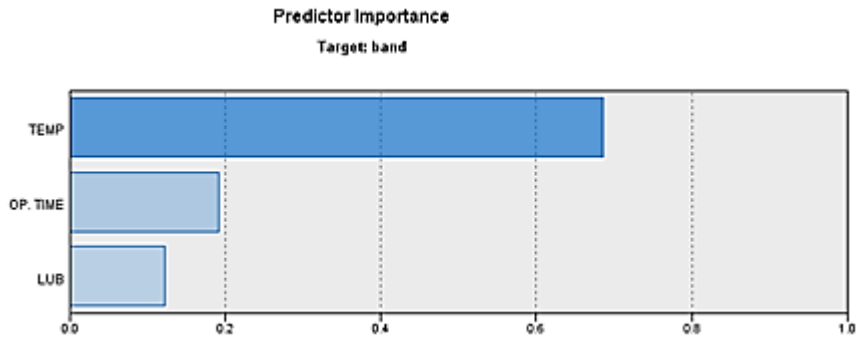


Fig. 9. The predictor importance levels obtained from the C5 decision tree.

Based on the results obtained, the extracted rules from the decision tree are as follows:

I. If the temperature is less than or equal to 19 degrees Celsius and the operation time is less than or equal to 165 hours, the variable 'band' is predicted to be 1. It implies that the number of failures per month will be less than 3.
II. If the temperature is less than or equal to 19 degrees Celsius and the operation time is greater than or equal to 165 hours, the variable 'band' is predicted to be 2. It implies that the number of failures per month will be between 3 and 5.
III. If the temperature is greater than 19 degrees Celsius and the oiling frequency is greater than 5 times, the variable 'band' is predicted to be 3. It implies that the number of failures per month will be between 5 and 8.
IV. If the temperature is greater than 19 degrees Celsius and the oiling frequency is less than or equal to 5 times, the variable 'band' is predicted to be 4. It implies that the number of failures per month will be more than 8.
V. In other cases, option 2 for the variable 'band' is predicted, meaning that the number of failures per month will be between 3 and 5.

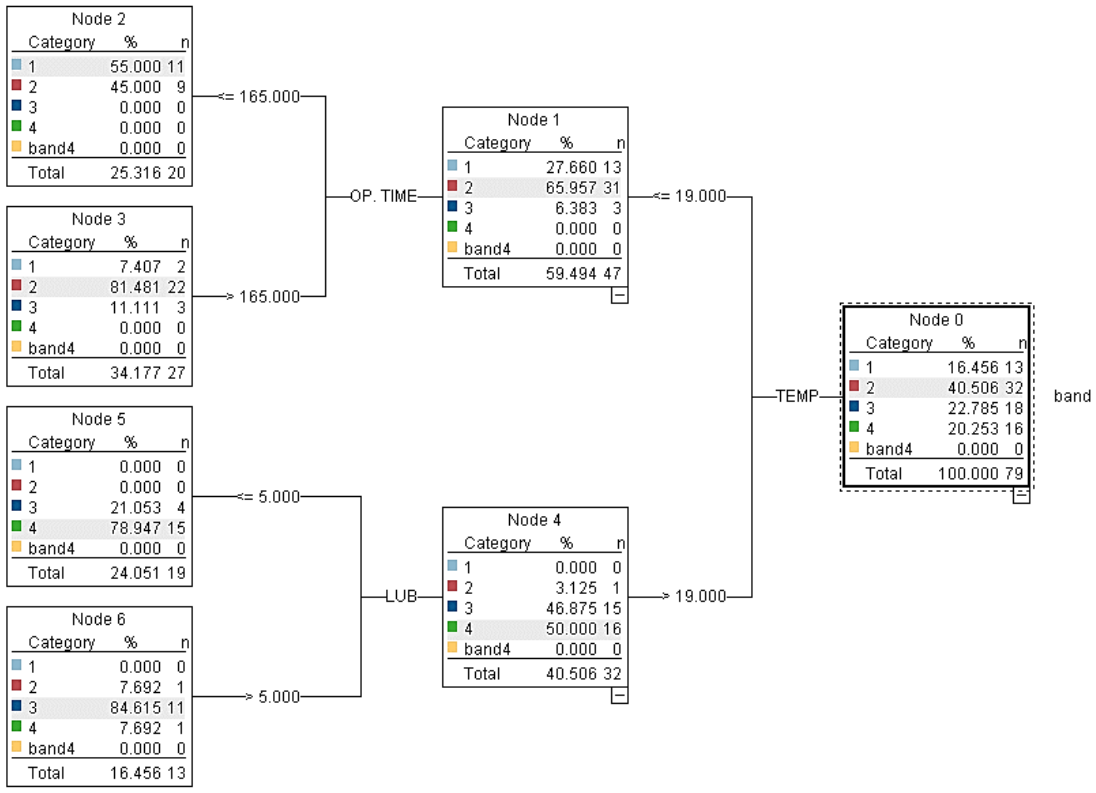


Fig. 10. The regression tree.

7 | Evaluation of the Model Accuracy

The output shown in *Fig. 5* was utilized to assess the accuracy of the model created using the analysis node in the modeler software. Based on the obtained results, the model's accuracy on the training data is approximately 77%, and on the testing data, it is around 76%. It means that out of the 79 testing instances used, the model correctly predicted 61 and made 18 incorrect predictions. In the case of the testing data, out of 85 instances, the model made 65 correct predictions and 20 incorrect predictions.

Table 5. Results of running the analysis node.

Partition	1_Training		2_Testing	
Correct	61	77.22%	65	76.47%
Wrong	18	22.78%	20	23.53%
Total	79		85	

8 | Conclusion

This research adds valuable insights to the burgeoning realm of ML applications for predicting water pump failures. By emphasizing often-neglected variables, the study broadens the horizons of predictive models. The comparison of multiple regression and decision tree cart methodologies underscores their efficacy in tackling the intricacies of failure prediction. As a next step, future research could delve into additional variables, refining predictive models to elevate further the reliability and accuracy of failure predictions in water pump systems. The findings not only contribute to the academic discourse but also hold practical implications for improving the maintenance and performance of water pump systems.

References

[1] Ikramov, N., Kan, E., Mirzoev, M., & Majidov, T. (2019). Effect of parallel connection of pumping units on operating costs of pumping station. In *E3S Web of Conferences* (Vol. 97, p. 05014). EDP Sciences.

[2] Ergashev, R., Bekchanov, F., Akmalov, S., Shodiev, B., & Kholbutaev, B. (2020). New methods for geoinformation systems of tests and analysis of causes of failure elements of pumping stations. *IOP conference series: materials science and engineering* (Vol. 883, p. 12015). IOP Publishing. DOI: 10.1088/1757-899X/883/1/012015

[3] Jacobs, J. A., Mathews, M. J., & Kleingeld, M. (2018). Failure prediction of mine de-watering pumps. *Journal of failure analysis and prevention*, *18*(4), 927–938.

[4] Mohammed, A. (2023). Data driven-based model for predicting pump failures in the oil and gas industry. *Engineering failure analysis*, *145*, 107019. DOI:10.1016/j.engfailanal.2022.107019

[5] Trstenjak, B., Palasek, B., & Trstenjak, J. (2019). A decision support system for the prediction of wastewater pumping station failures based on CBR continuous learning model. *Engineering, technology and applied science research*, *9*(5), 4745–4749. DOI:10.48084/etasr.3031

[6] Afshar-Nadjafi, B., Pourbakhsh, H., Mirhabibi, M., Khodaei, H., Ghodami, B., Sadighi, F., & Azizi, S. (2019). Economic production quantity model with backorders and items with imperfect/perfect quality options. *Journal of applied research and technology*, *17*(4), 250–257. DOI:10.22201/icat.16656423.2019.17.4.794

[7] Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., ... & Herrera, F. (2023). Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. *Information fusion*, *99*, 101805. DOI:10.1016/j.inffus.2023.101805

[8] Azizi, S., & Mohammadi, M. (2023). Strategy selection for multi-objective redundancy allocation problem in a k-out-of-n system considering the mean time to failure. *Opsearch*, *60*(2), 1021–1044. DOI:10.1007/s12597-023-00635-2

[9] Issah, I., Appiah, O., Appiahene, P., & Inusah, F. (2023). A systematic review of the literature on machine learning application of determining the attributes influencing academic performance. *Decision analytics journal*, *7*, 100204. DOI:10.1016/j.dajour.2023.100204

[10] Rasi Nojehdehi, R., Bagherzadeh Valami, H., & Najafi, S. E. (2023). Classifications of linking activities based on their inefficiencies in network DEA. *International journal of research in industrial engineering*, *12*(2), 165–176.

[11] Rasinojehdehi, R., & Valami, H. B. (2023). A comprehensive neutrosophic model for evaluating the efficiency of airlines based on SBM model of network DEA. *Decision making: applications in management and engineering*, *6*(2), 880–906. DOI:10.31181/dma622023729

[12] Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., ... & Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, *14*(1), 1–37. DOI:10.1007/s10115-007-0114-2

[13] Kurani, A., Doshi, P., Vakharia, A., & Shah, M. (2023). A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting. *Annals of data science*, *10*(1), 183–208. DOI:10.1007/s40745-021-00344-x

[14] Nojehdehi, R. R., Maleki, P., Abianeh, M., & Valami, H. B. (2012). A geometrical approach for fuzzy production possibility set in data envelopment analysis (DEA) with fuzzy input-output levels. *African journal of business management*, *6*(7), 2738–2745.

[15] Reynara, F. J., Carolina, S., & Simbolon, I. N. (2022). The comparison of C4.5 and CART (classification and regression tree) algorithm in classification of occupation for fresh graduate. *IConVET 2021: proceedings of the 4th international conference on vocational education and technology* (p. 13). European Alliance for Innovation. DOI: 10.4108/eai.27-11-2021.2315527

[16] Bansal, M., Goyal, A., & Choudhary, A. (2022). A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. *Decision analytics journal*, *3*, 100071. DOI:10.1016/j.dajour.2022.100071

[17] Hasri, C. F., & Alita, D. (2022). Penerapan metode naïve bayes classifier dan support vector machine pada analisis sentimen terhadap dampak virus corona di twitter. *Jurnal informatika dan rekayasa perangkat lunak*, *3*(2), 145–160. DOI:10.33365/jatika.v3i2.2026

[18] Herrera, G., & Morillo, P. (2022). Benchmarking of supervised machine learning algorithms in the early failure prediction of a water pumping system. *Communication, smart technologies and innovation for society: proceedings of CITIS 2021* (pp. 535–546). Springer. https://doi.org/10.1007/978-981-16-4126-8_48

[19] Velasco Robles, A. (2022). *A machine learning approach to predict pipe failures in water distribution networks* (Ph.D Thesis, Universidad de Sevilla). https://dialnet.unirioja.es/servlet/dctes?codigo=305976

[20] Sunal, C. E., Dyo, V., & Velisavljevic, V. (2022). Review of machine learning based fault detection for centrifugal pump induction motors. *IEEE access*, *10*, 71344–71355. DOI:10.1109/ACCESS.2022.3187718

[21] Eiben, A. E., Berends, T., & Mosch, T. (2022). *Predictive maintenance for sewage pumping stations using machine learning* (Ph.D Thesis, Vrije Universiteit Amsterdam). https://vu-business-analytics.github.io/internship-office/reports/report-internnn.pdf

[22] Kreuzberger, D., Kuhl, N., & Hirschl, S. (2023). Machine learning operations (mlops): overview, definition, and architecture. *IEEE access*, *11*, 31866–31879. DOI:10.1109/ACCESS.2023.3262138

[23] Eberly, L. E. (2007). Multiple linear regression. *Topics in biostatistics*, 165–187. https://doi.org/10.1007/978-1-59745-530-5_9

[24] Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of applied science and technology trends*, *2*(01), 20–28. DOI:10.38094/jastt20165